ArrayMiner™
ClassMarker Tutorial

Version 3.0

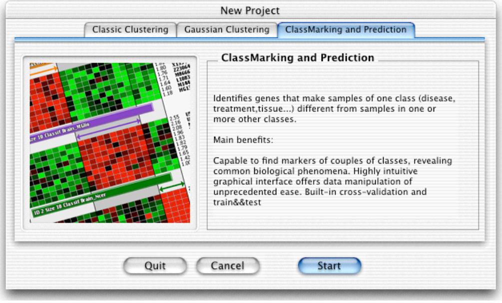
www.optimaldesign.com

1. Introduction

ArrayMiner™ ClassMarker is a class marking and prediction tool that enables users to understand what differentiates samples in various classes (diseases, treatments…) at molecular (gene) level, ie to identify markers of the classes. These markers can then be used to predict the class of unknown samples.

2. Running a cross validation

Open the file named "DataSet_A_AllClasses.csv" using the open file menu or the tool bar button.



Identifies genes that make samples of one class (disease, treatment,tissue...) different from samples in one or more other classes.

Main benefits:

Capable to find markers of couples of classes, revealing common biological phenomena. Highly intuitive graphical interface offers data manipulation of unprecedented ease. Built-in cross-validation and train&&test

Figure 1

Choose the Class Marker option by clicking on the third tab of the previous figure. Click on the start button. The following figure will appear showing a heat map view of the data.

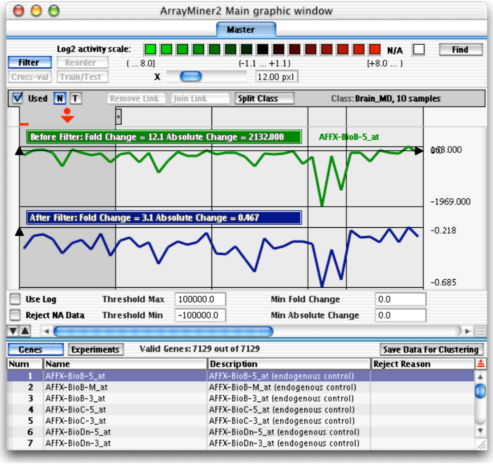
Class: Brain_MD, 10 samples

Valid Genes: 7129 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	
7	AFFX-BioDn-3_at	AFFX-BioDn-3_at (endogenous control)	

Figure 2

The next step is to specify filter values for the raw data. Click on the filter button and the following figure will appear.



☐ Use Log — Threshold Max: 100000.0 — Min Fold Change: 0.0

☐ Reject NA Data — Threshold Min: -100000.0 — Min Absolute Change: 0.0

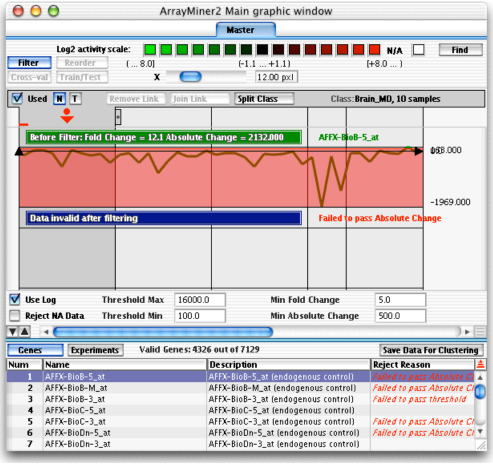
Valid Genes: 7129 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	
7	AFFX-BioDn-3_at	AFFX-BioDn-3_at (endogenous control)	

Figure 3

Change the values of the filter to match the following figure. The values are located in the center of the window. There are six of them :

- **Use log** [Set it to Yes]: Apply a log transformation to all measurements
- **Reject N/A Data** [Set it to No]: Reject a gene with missing values
- **Threshold max** [Set it to 16000]: The maximum value of the threshold
- **Threshold min** [Set it to 100]: The minimum value of the threshold
- **Min fold change** [Set it to 5]: The minimum value of the ratio between the maximum and the minimum measurements of a gene to keep it in the analysis
- **Min absolute fold change** [Set it to 500]: The minimum absolute difference beween the maximum and the minimum measurements of a gene to keep it in the analysis



☑ Use Log — Threshold Max: 16000.0 — Min Fold Change: 5.0

☐ Reject NA Data — Threshold Min: 100.0 — Min Absolute Change: 500.0

Valid Genes: 4326 out of 7129

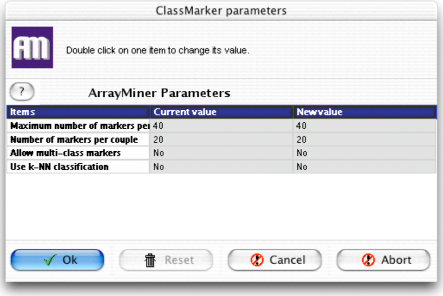
Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	*Failed to pass Absolute C*
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	*Failed to pass Absolute C*
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	*Failed to pass threshold*
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	*Failed to pass Absolute C*
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	*Failed to pass Absolute C*
7	AFFX-BioDn-3_at	AFFX-BioDn-3_at (endogenous control)	

Figure 4

You can view how the filter transforms each gene by clicking on the appropriate line in the bottom table. There should be now 4326 valid genes out of 7129. (Information drawn in the center of the screen).

Quit the filter mode by clicking one more time on the “Filter” button.

Click on the “Cross-Val” button to start analysis. A parameter box will appear as shown in the following figure.



Double click on one item to change its value.

ArrayMiner Parameters

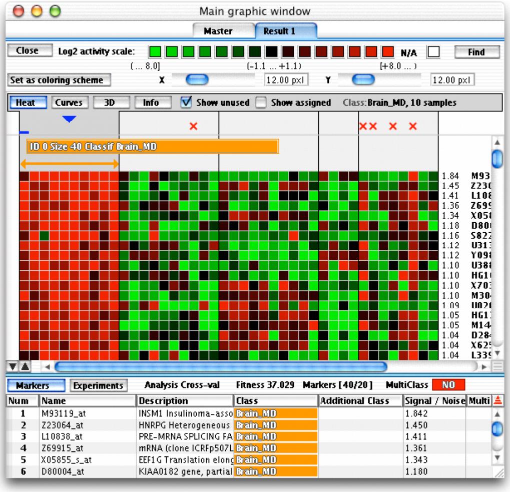
Items	Current value	Newvalue
Maximum number of markers pe	40	40
Number of markers per couple	20	20
Allow multi-class markers	No	No
Use k-NN classification	No	No

Figure 5

There are five options available. Change them to match the values of the previous figure.(Double click on a value to change it)

- **Maximum number of markers** [Set it to 40]: This will limit to this value the number of markers used to differenciate a class.
- **Number of markers per couple** [Set it to 20]: This will limit to this value the number of additional markers per couple of classes.
- **Allow multi-class markers** [Set it to No]: This will enable/disable the multiclass analysis of the data.
- **Use k-NN classification** [Set it to No]: When set to yes, the vote is computed using the k-Nearest-Neighbour algorithm.

Click on OK. After a short computation time a new tab will appear in the main window named “Result 1”.



X 12.00 pxl Y 12.00 pxl

☑ Show unused ☐ Show assigned Class:Brain_MD, 10 samples

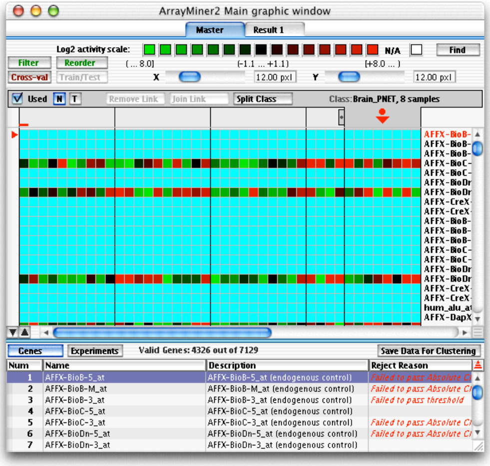
Analysis Cross-val Fitness 37.029 Markers [40/20] MultiClass NO

Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	M93119_at	INSM1 Insulinoma-asso	Brain_MD		1.842	
2	Z23064_at	HNRPG Heterogeneous	Brain_MD		1.450	
3	L10838_at	PRE-MRNA SPLICING FA	Brain_MD		1.411	
4	Z69915_at	mRNA (clone ICRFp507L	Brain_MD		1.361	
5	X05855_s_at	EEF1G Translation elong	Brain_MD		1.343	
6	D80004_at	KIAA0182 gene, partial	Brain_MD		1.180	

Figure 6

This tab is the result of the previous analysis, you can change its name by double-clicking the name. As depicted in the previous figure, red crosses show the cross validation failures for each sample. By scrolling vertically, you can scan the group of markers discriminating each class. You can switch to different visualization mode by clicking on the “Heat”, “Curves”, ”3D”, and “Info” buttons.

As shown in the previous figure, there are four failures in the last class. You can easily inform the algorithm to ignore the class in the computation. Return to the master tab by clicking on it and select the last class named “Brain_PNET” by clicking on it in the class selection area located just above the heat map as shown in the next figure.



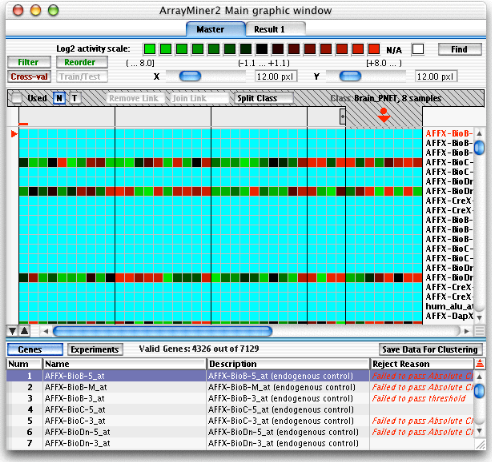
Class: Brain_PNET, 8 samples

Valid Genes: 4326 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	Failed to pass Absolute Cl
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	Failed to pass Absolute Cl
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	Failed to pass threshold
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	Failed to pass Absolute Cl
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	Failed to pass Absolute Cl
7	AFFX-BioDn-3_at	AFFX-BioDn-3_at (endogenous control)	

Figure 7

Click on the checkbox “Used” to uncheck it. This will change the class status to unused. The class is now dashed as shown in the following figure.



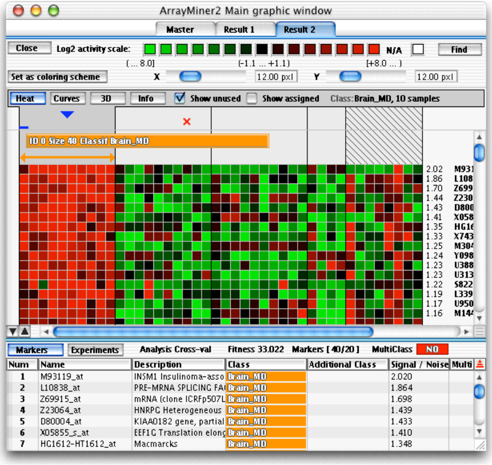
Class: Brain_PNET, 8 samples

Valid Genes: 4326 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	Failed to pass Absolute Cl
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	Failed to pass Absolute Cl
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	Failed to pass threshold
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	Failed to pass Absolute Cl
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	Failed to pass Absolute Cl
7	AFFX-BioDn-3_at	AFFX-BioDn-3_at (endogenous control)	

Figure 8

Click one more time on the “Cross-Val” to start a new analysis, click OK in the parameters box and wait for the analysis to complete.



☑ Show unused ☐ Show assigned Class: Brain_MD, 10 samples

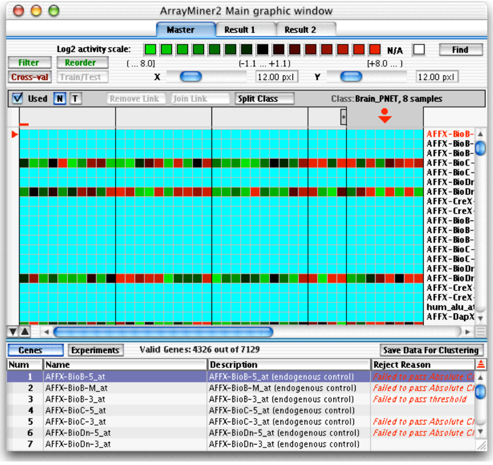
Analysis Cross-val Fitness 33.022 Markers [40/20] MultiClass NO

Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	M93119_at	INSM1 Insulinoma-asso	Brain_MD		2.020	
2	L10838_at	PRE-MRNA SPLICING FA	Brain_MD		1.864	
3	Z69915_at	mRNA (clone ICRFp507L	Brain_MD		1.698	
4	Z23064_at	HNRPG Heterogeneous	Brain_MD		1.439	
5	D80004_at	KIAA0182 gene, partial	Brain_MD		1.433	
6	X05855_s_at	EEF1G Translation elong	Brain_MD		1.410	
7	HG1612-HT1612_at	Macmarcks	Brain_MD		1.348	

Figure 9

There is now only one cross validation failure as shown in the previous figure (There is only one red cross). You can also try to keep the last class (Brain_PNET) and use the multi-class possibilities of the algorithm.

To do so, return to the master tab by clicking on it. Select the last class, check the "Used" check box, you should obtain the following figure.



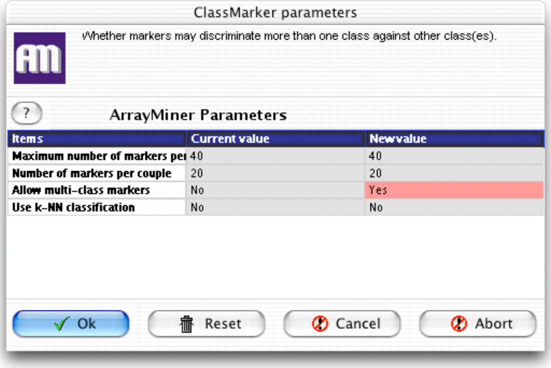
☑ Used Class: Brain_PNET, 8 samples

Valid Genes: 4326 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	Failed to pass Absolute Cl
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	Failed to pass Absolute Cl
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	Failed to pass threshold
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	Failed to pass Absolute Cl
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	Failed to pass Absolute Cl
7	AFFX-BioDn-3_at	AFFX-BioDn-3_at (endogenous control)	

Figure 10

Click one more time on the "Cross-Val" button. When the parameters box appears, change the multi-class option to yes as shown in the following figure.

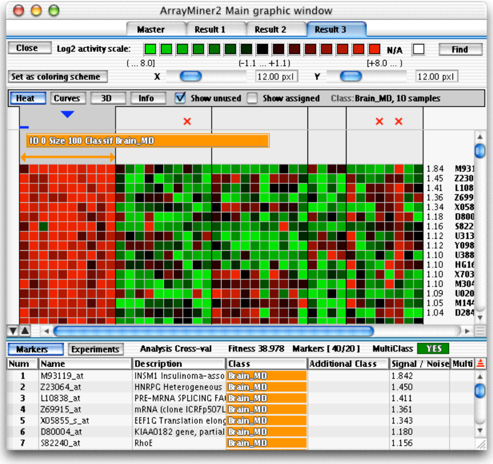


Whether markers may discriminate more than one class against other class(es).

Items	Current value	New value
Maximum number of markers pe	40	40
Number of markers per couple	20	20
Allow multi-class markers	No	Yes
Use k-NN classification	No	No

Figure 11

At the end of the analysis, you should receive the following result.



Analysis Cross-val Fitness 38.978 Markers [40/20] MultiClass YES

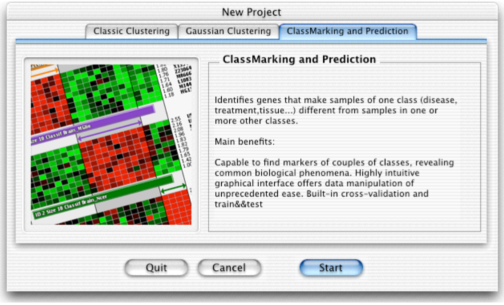
Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	M93119_at	INSM1 Insulinoma-asso	Brain_MD		1.842	
2	Z23064_at	HNRPC Heterogeneous	Brain_MD		1.450	
3	L10838_at	PRE-MRNA SPLICING FA	Brain_MD		1.411	
4	Z69915_at	mRNA (clone ICRFp507L	Brain_MD		1.361	
5	X05855_s_at	EEF1G Translation elong	Brain_MD		1.343	
6	D80004_at	KIAA0182 gene, partial	Brain_MD		1.180	
7	S82240_at	RhoE	Brain_MD		1.156	

Figure 12

There are three cross-validation failures, but thanks to the multi-class algorithm, we now considered the full dataset.

3. Running a Train and Test analysis

Open the file named “ALL_AML_TrainAndTest_AllClass.csv” using the open file menu or the tool bar button.



ClassMarking and Prediction

Identifies genes that make samples of one class (disease, treatment,tissue...) different from samples in one or more other classes.

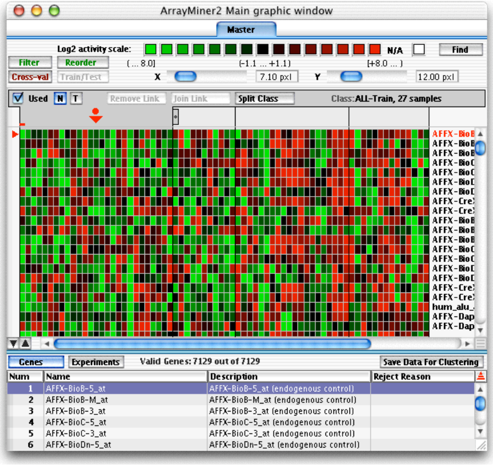
Main benefits:

Capable to find markers of couples of classes, revealing common biological phenomena. Highly intuitive graphical interface offers data manipulation of unprecedented ease. Built-in cross-validation and train&&test

Figure 13

Choose the Class Marker option by clicking on the third tab of the previous figure. Click on the start button.

You should view the following figure.



X: 7.10 pxl Y: 12.00 pxl

Class:ALL-Train, 27 samples

Valid Genes: 7129 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	

Figure 14

As for the “Cross-Validation” analysis, you have to first define the filter values for your data by setting them in the filter mode. Specify the same values as in the previous tutorial. Note that the filter default values are set to the last used ones, so the previous figure may have some genes filtered out (light cyan colored) depending on the filter values used the last time.

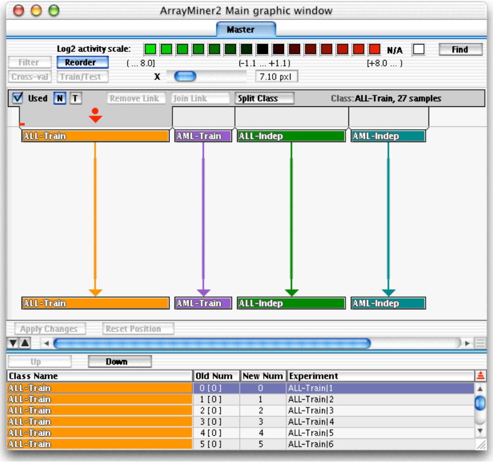
The data are separated into four classes.

- ALL-Train
- AML-Train
- ALL-Indep
- AML-Indep

To facilitate the viewing of the data, it could be interesting to switch the position of the AML-Train with the ALL-Indep to obtain the classes in the following order :

- ALL-Train
- ALL-Indep
- AML-Train
- AML-Indep

To obtain this order, switch to the reorder mode by clicking on the “Reorder” button located in the top left part of the window. You will obtain the following figure.

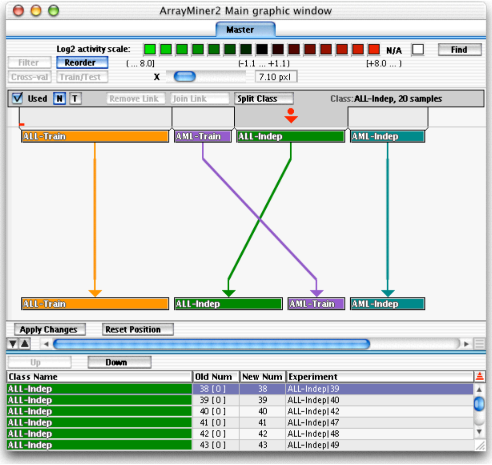


Class: ALL-Train, 27 samples

Class Name	Old Num	New Num	Experiment
ALL-Train	0 [0]	0	ALL-Train|1
ALL-Train	1 [0]	1	ALL-Train|2
ALL-Train	2 [0]	2	ALL-Train|3
ALL-Train	3 [0]	3	ALL-Train|4
ALL-Train	4 [0]	4	ALL-Train|5
ALL-Train	5 [0]	5	ALL-Train|6

Figure 15

To invert the position of the two classes, you can for example, drag the “ALL-Indep” class (Green box) to its new position as shown in the following picture.

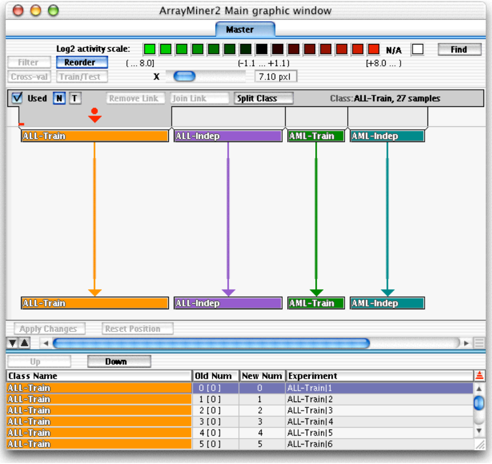


Class: ALL-Indep, 20 samples

Class Name	Old Num	New Num	Experiment
ALL-Indep	38 [0]	38	ALL-Indep|39
ALL-Indep	39 [0]	39	ALL-Indep|40
ALL-Indep	40 [0]	40	ALL-Indep|42
ALL-Indep	41 [0]	41	ALL-Indep|47
ALL-Indep	42 [0]	42	ALL-Indep|48
ALL-Indep	43 [0]	43	ALL-Indep|49

Figure 16

You can now click on the “Apply changes” button to validate the modifications. You will obtain the following figure.



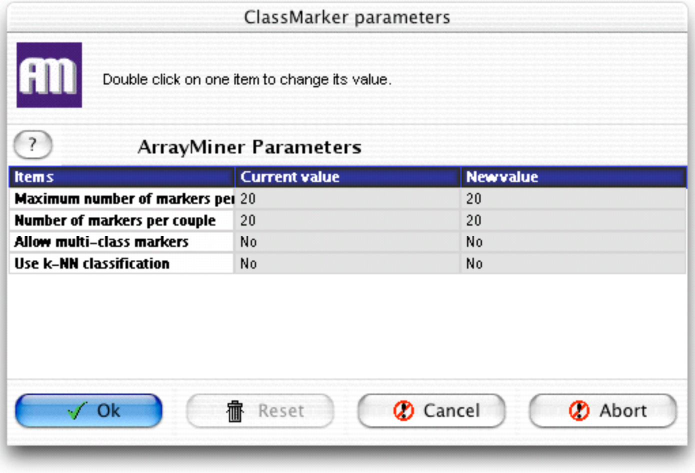
Class Name	Old Num	New Num	Experiment
ALL-Train	0 [0]	0	ALL-Train|1
ALL-Train	1 [0]	1	ALL-Train|2
ALL-Train	2 [0]	2	ALL-Train|3
ALL-Train	3 [0]	3	ALL-Train|4
ALL-Train	4 [0]	4	ALL-Train|5
ALL-Train	5 [0]	5	ALL-Train|6

Figure 17

Exit the reorder mode by clicking once more on the “Reorder” button. We now have the classes in this order:

- ALL-Train
- ALL-Indep
- AML-Train
- AML-Indep

We can try a cross-validation by clicking on the “Cross-Val” button. Choose the parameters in the Edit Parameters Box to match the following figure.

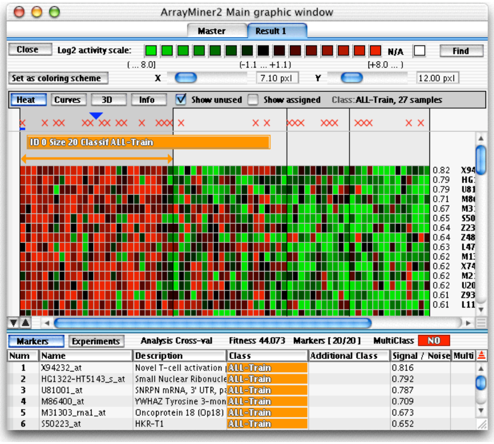


Double click on one item to change its value.

Items	Current value	Newvalue
Maximum number of markers pei	20	20
Number of markers per couple	20	20
Allow multi-class markers	No	No
Use k-NN classification	No	No

Figure 18

Click on “Ok”, you will then obtain the following result.



☑ Show unused ☐ Show assigned Class: ALL-Train, 27 samples

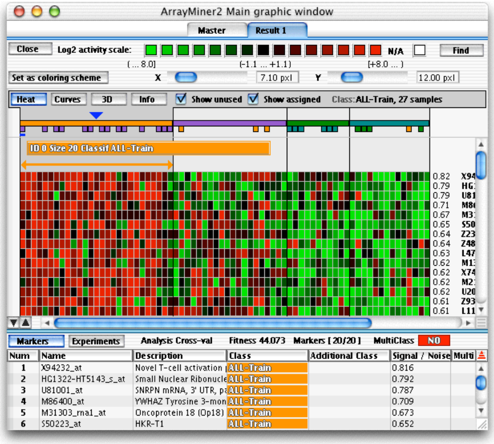
Markers | Experiments | Analysis Cross-val | Fitness 44.073 | Markers [20/20] | MultiClass NO

Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	X94232_at	Novel T-cell activation	ALL-Train		0.816	
2	HG1322-HT5143_s_at	Small Nuclear Ribonucle	ALL-Train		0.792	
3	U81001_at	SNRPN mRNA, 3' UTR, pa	ALL-Train		0.787	
4	M86400_at	YWHAZ Tyrosine 3-mon	ALL-Train		0.709	
5	M31303_rna1_at	Oncoprotein 18 (Op18)	ALL-Train		0.673	
6	S50223_at	HKR-T1	ALL-Train		0.652	

Figure 19

As shown in the previous figure, there are many cross-validation failures, but in fact this is normal because we ask the algorithm to differentiate four classes, where there are obviously only two (ALL and AML).

You can visualize how the classes are mixed by clicking on the “Show Assigned” check box, to realize that ClassMarker mixed the “ALL-Indep” set with the “ALL-Train” set and the “AML-Indep” set with the “AML-Train” set, as shown in the following figure.



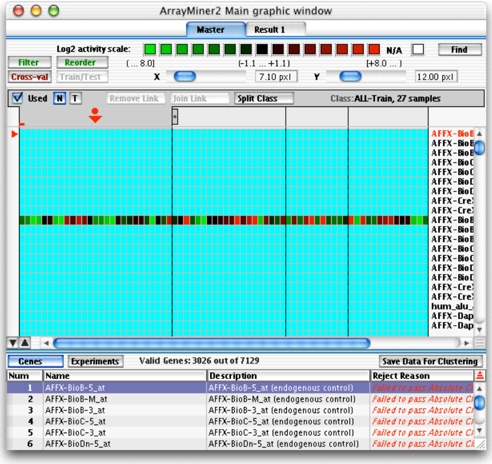
☑ Show unused ☑ Show assigned Class: ALL-Train, 27 samples

Markers | Experiments | Analysis Cross-val | Fitness 44.073 | Markers [20/20] | MultiClass NO

Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	X94232_at	Novel T-cell activation	ALL-Train		0.816	
2	HG1322-HT5143_s_at	Small Nuclear Ribonucle	ALL-Train		0.792	
3	U81001_at	SNRPN mRNA, 3' UTR, pa	ALL-Train		0.787	
4	M86400_at	YWHAZ Tyrosine 3-mon	ALL-Train		0.709	
5	M31303_rna1_at	Oncoprotein 18 (Op18)	ALL-Train		0.673	
6	S50223_at	HKR-T1	ALL-Train		0.652	

Figure 20

We will now perform a “Train and Test” analysis. Return to the master tab by clicking on its name. Select the first class “ALL-Train” as shown in the following figure. You can also select a class with help of the “Left” and “Right” keys.

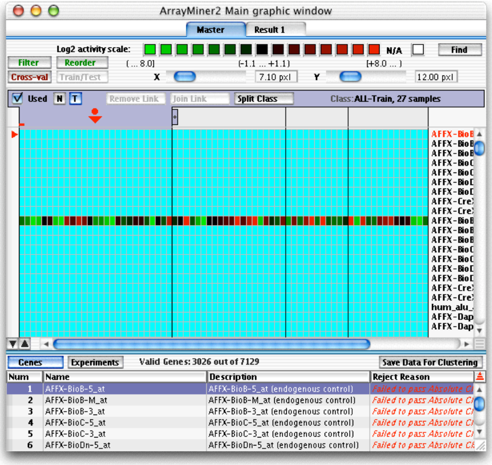


Valid Genes: 3026 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	Failed to pass Absolute C
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	Failed to pass Absolute C
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	Failed to pass Absolute C
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	Failed to pass Absolute C
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	Failed to pass Absolute C
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	Failed to pass Absolute C

Figure 21

Click on the “T” button located near the “Used” check button. The whole class will be colored in light blue, meaning that it becomes a “Train” class as shown in the following figure.

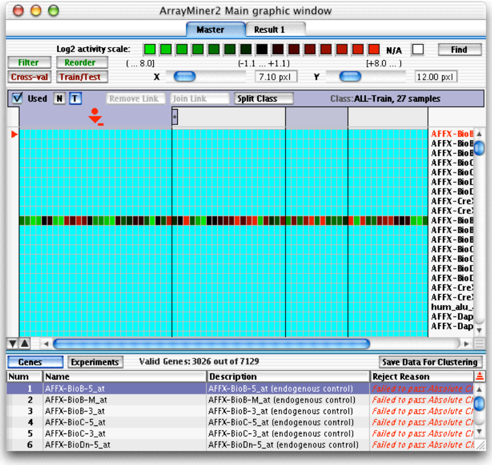


Valid Genes: 3026 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	Failed to pass Absolute C
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	Failed to pass Absolute C
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	Failed to pass Absolute C
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	Failed to pass Absolute C
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	Failed to pass Absolute C
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	Failed to pass Absolute C

Figure 22

Repeat the previous step for the third class named “AML-Train”. You will obtain the following figure.

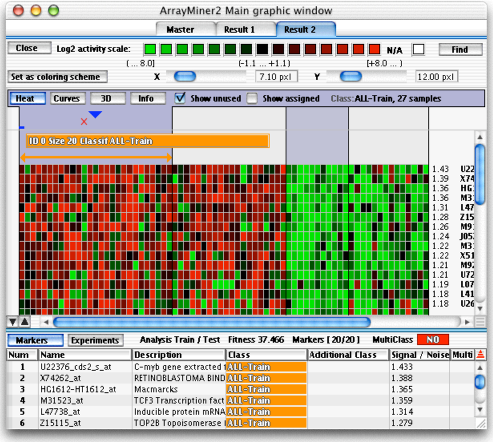


Valid Genes: 3026 out of 7129

Num	Name	Description	Reject Reason
1	AFFX-BioB-5_at	AFFX-BioB-5_at (endogenous control)	Failed to pass Absolute Cl
2	AFFX-BioB-M_at	AFFX-BioB-M_at (endogenous control)	Failed to pass Absolute Cl
3	AFFX-BioB-3_at	AFFX-BioB-3_at (endogenous control)	Failed to pass Absolute Cl
4	AFFX-BioC-5_at	AFFX-BioC-5_at (endogenous control)	Failed to pass Absolute Cl
5	AFFX-BioC-3_at	AFFX-BioC-3_at (endogenous control)	Failed to pass Absolute Cl
6	AFFX-BioDn-5_at	AFFX-BioDn-5_at (endogenous control)	Failed to pass Absolute Cl

Figure 23

Click on the “Train/Test button to start the analysis. Reply “Yes” to the question “Cross-Validate the train set”. You will obtain the following result.

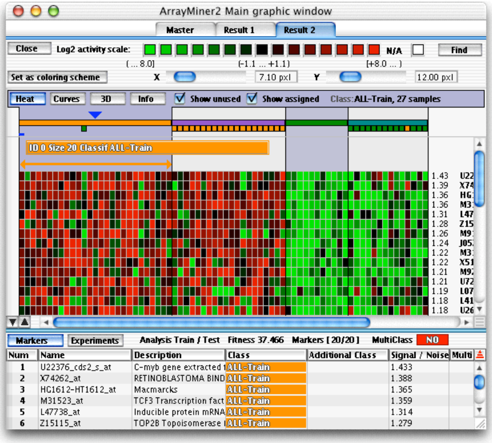


Analysis Train / Test Fitness 37.466 Markers [20/20] MultiClass NO

Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	U22376_cds2_s_at	C-myb gene extracted f	ALL-Train		1.433	
2	X74262_at	RETINOBLASTOMA BIND	ALL-Train		1.388	
3	HG1612-HT1612_at	Macmarcks	ALL-Train		1.365	
4	M31523_at	TCF3 Transcription fact	ALL-Train		1.359	
5	L47738_at	Inducible protein mRNA	ALL-Train		1.314	
6	Z15115_at	TOP2B Topoisomerase I	ALL-Train		1.279	

Figure 24

The markers are obtained by using only the two blue classes (Train classes). There was only one cross-validation failure, in the first class. The samples in the other two classes (gray) are classified by voting. To view their corresponding assigned class, simply click on the “Show Assigned” check box and you will obtain the next figure.



X 7.10 pxl Y 12.00 pxl

☑ Show unused ☑ Show assigned Class: ALL-Train, 27 samples

Analysis Train / Test Fitness 37.466 Markers [20/20] MultiClass NO

Num	Name	Description	Class	Additional Class	Signal / Noise	Multi
1	U22376_cds2_s_at	C-myb gene extracted f	ALL-Train		1.433	
2	X74262_at	RETINOBLASTOMA BIND	ALL-Train		1.388	
3	HG1612-HT1612_at	Macmarcks	ALL-Train		1.365	
4	M31523_at	TCF3 Transcription fact	ALL-Train		1.359	
5	L47738_at	Inducible protein mRNA	ALL-Train		1.314	
6	Z15115_at	TOP2B Topoisomerase I	ALL-Train		1.279	

Figure 25

There is only one misclassified sample in the test classes, as shown in the previous figure.